A Simulation Study of Confounding in Generalized Linear Models for Air Pollution Epidemiology

Colin Chen, David P. Chock, and Sandra L. Winkler

Ford Research Laboratory, Dearborn, MI 48121 USA

Confounding between the model covariates and causal variables (which may or may not be included as model covariates) is a well-known problem in regression models used in air pollution epidemiology. This problem is usually acknowledged but hardly ever investigated, especially in the context of generalized linear models. Using synthetic data sets, the present study shows how model overfit, underfit, and misfit in the presence of correlated causal variables in a Poisson regression model affect the estimated coefficients of the covariates and their confidence levels. The study also shows how this effect changes with the ranges of the covariates and the sample size. There is qualitative agreement between these study results and the corresponding expressions in the large-sample limit for the ordinary linear models. Confounding of covariates in an overfitted model (with covariates encompassing more than just the causal variables) does not bias the estimated coefficients but reduces their significance. The effect of model underfit (with some causal variables excluded as covariates) or misfit (with covariates encompassing only noncausal variables), on the other hand, leads to not only erroneous estimated coefficients, but a misguided confidence, represented by large *t*-values, that the estimated coefficients are significant. The results of this study indicate that models which use only one or two air quality variables, such as particulate matter ≤10 μm and sulfur dioxide, are probably unreliable, and that models containing several correlated and toxic or potentially toxic air quality variables should also be investigated in order to minimize the situation of model underfit or misfit. *Key words*: air pollution, confounding effects, ecologic time–series studies, epidemiological studies, generalized linear models, model misfit, Poisson regression, simulation. *Environ Health Perspect* 107:217–222 (1999). [Online 8 February 1999]
http://ehpnet1.niehs.nih.gov/docs/1999/107p217-222chen/abstract.html

The EPA (*1*) recently promulgated the National Ambient Air Quality Standards for the mass concentrations of particulate matter ≤10 μm (PM_{10}) and ≤2.5 μm ($PM_{2.5}$). The key rationale for these standards came from the epidemiological studies in the past few years associating particulate air pollution [represented in these studies primarily by the ambient concentrations of either the total suspended particulate (TSP) or PM_{10}] with daily mortality and morbidity. Both the mortality and morbidity studies are almost exclusively ecological time–series studies regressing the daily events of mortality or morbidity against the ambient air quality for given urban areas (*2*). However, for many urban areas where the same or similar data sets were reanalyzed, different or contradictory conclusions often resulted (*3–14*). This fact highlights the difficulty of establishing a causal relation between ambient PM concentrations at their present levels and a given health endpoint through regression models alone.

In most existing Poisson regression models for countable data such as mortality and morbidity, in addition to the inclusion of some weather parameters, ambient concentrations of some air pollutants, often only one (PM_{10} or TSP) or two (PM_{10} or TSP, and SO_2) so far, were included as covariates. Models were then selected based on some goodness-of-fit criteria. However, different choices of covariates and different formulations of the model led to different selected models with different accompanying conclusions, and it is often impossible to ascertain which model is more correct or reliable. One major issue among many is the issue of confounding or collinearity, which we take to mean the presence of significant correlation between a covariate in a regression model and another covariate that may or may not be causal, or a causal variable that may or may not be a covariate in the model. This issue is always acknowledged but almost never investigated. In fact, after acknowledgment of a potential confounding problem, most researchers went on to draw conclusions based on the significance of the estimated coefficients associated with given covariates, often oblivious of or ignoring the fact that confounding can invalidate the conclusions altogether. It is also not often appreciated that uncovering similar regression results for many areas does not necessarily reduce or remove the problem of confounding because the same confounding problem may occur in many areas. However, in ascertaining if PM among many pollutants is indeed a causal agent of daily mortality or morbidity, the problem of confounding cannot be ignored, especially when correlation between ambient concentrations of different air pollutants can be large. Table 1 shows an example of typical correlation coefficients among the air quality variables and meteorological variables that have been considered as covariates in regression studies. The data are from Pittsburgh (Allegheny County), Pennsylvania, for the summer and winter seasons from 1989 to 1991. The pollutants are all daily maximum 1-hr concentrations, with the exception of PM_{10}, which is a daily-averaged concentration. The meteorological variables are for the hour of the day when the dry-bulb temperature reaches the maximum. Table 1 shows that relatively high correlation can exist between different air quality and meteorological variables, and that the correlation is seasonally dependent. For example, the correlation between O_3 and PM_{10} is high in the summer but very low in the winter. On the other hand, the correlation between CO and PM_{10} is moderately high in the summer and quite high in the winter. Also, both O_3 and PM_{10} are well correlated with dry-bulb temperature in the summer but less well correlated in the winter.

Depending on the choice of covariates in a model, different types of model biases can occur. They include underfitting (with some causal variables being excluded as covariates), overfitting (with covariates including more than just the causal variables) and misfitting (with all covariates being noncausal variables). The effect of such a model bias on the estimated coefficients of the chosen covariates in an ordinary least squares (OLS) linear model was briefly described by Seber (*15*). Chen and Zhang (*16*) have provided the means and variances of the estimated coefficients in the large-sample limit (number of observations being much greater than one) for a biased OLS model. However, no closed form is available to describe the impact of model bias due to underfitting, overfitting, and misfitting of covariates on the estimated coefficients of a generalized linear model (GLM) (*17*). This knowledge has great theoretical and practical importance in view of the role of Poisson regression in the PM epidemiological studies. It is difficult to study the confounding effects because no one knows what a correct model

Address correspondence to D.P. Chock, Ford Research Laboratory, P.O. Box 2053, MD-3083, Dearborn, MI 48121-2053 USA. C. Chen is currently at the SAS Institute, Cary,NC 27513 USA.
Received 29 January 1998; accepted 21 September 1998.

Table 1. Correlation coefficients among the ambient pollutant concentrations and weather parameters for Pittsburgh, Pennsylvania, 1989–1991 (summer/winter in the upper right/lower left, relative to the diagonal)

	CO	NO_2	O_3	SO_2	PM_{10}	TempDry	DewPt
CO	1.00	0.61	0.28	0.44	0.44	0.15	0.03
NO_2	0.76	1.00	0.56	0.46	0.62	0.46	0.03
O_3	-0.05	0.11	1.00	0.30	0.66	0.70	0.15
SO_2	0.40	0.41	-0.09	1.00	0.43	0.17	-0.04
PM_{10}	0.72	0.73	-0.06	0.54	1.00	0.58	0.39
TempDry	0.31	0.32	0.34	0.09	0.30	1.00	0.46
DewPt	0.15	0.00	0.07	-0.09	0.04	0.74	1.00

Abbreviations: PM_{10}, particulate matter ≤10 μm; TempDry, dry-bulb temperature; DewPt, dew point.

should be. However, synthetic data sets based on a known causal model can serve as a correct model for the study. The purpose of this paper is to investigate how model bias impacts the estimated coefficients and, more important, how covariate confounding, range of fluctuation of the covariates, and sample size affect this impact in a Poisson regression, assuming that we know *a priori* the exact causal relationship. Our approach is to construct synthetic data sets of a Poisson variate whose mean is determined by a known, exact linear model containing no more than two covariates that have a range of correlation coefficients between them. We then estimate the coefficients of the covariates for a Poisson regression model, biased or otherwise, applied to the synthetic data sets. In addition, the range of fluctuation of one of the covariates and the sample size will also be varied to see how they influence the *t*-values of the estimated coefficients. Serial correlation of the dependent and independent variables is not considered, as its inclusion would complicate the synthetic data set generation and would not add any new insight or alter the conclusion significantly. Seasonal cycles, long-term trends, and measurement error are not explicitly considered, as they are not a necessary component of confounding. However, they are relevant because they influence the correlation between the different air quality and meteorological variables, as well as the ranges of the variables. Therefore, they influence the extent of confounding. We also present the closed form results for the OLS models and describe the protocol for the construction of the synthetic data sets and the results of different Poisson regressions applied to the data sets.

Confounding Effects of Covariates in Ordinary Linear Models

Before the simulation study of GLM, we present a brief review of the results extracted from Chen and Zhang (*16*). Consider a linear regression model containing p covariates, x_j, (including $x_1 = 1$),

$$y = \sum_{j=1}^{p} \beta_j x_j + \varepsilon. \tag{1}$$

Here, y is the dependent variable; $X = (x_1, ..., x_p)^T$ is the covariate vector, with the corresponding expectation $E(X) = \mu$ and variance matrix $VAR(X) = \Sigma > 0$; β_j, $j = 1, ..., p$ are the coefficients of the covariates; ε is the random error, independent of X, with $E(\varepsilon) = 0$; $var(\varepsilon) = \sigma^2$. The OLS estimator of $\beta = (\beta_1, ..., \beta_p)^T$ is

$$\hat{\beta} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

where $\mathbf{Y} = (y_1, ..., y_n)^T$; $\mathbf{X} = (x_{ij})_{n\times p}$, n is the number of observations or the sample size.

Unbiased. If the data are actually generated from a model identical to Equation 1, we call the equation unbiased for the data. With these assumptions, we have in the large-sample limit ($n>>1$) the expectation and variance of the estimated coefficients,

$$E\left(\hat{\underline{\beta}}\right) = \underline{\beta}, \tag{3}$$

$$V\left(\hat{\underline{\beta}}\right) \sim \frac{\sigma^2}{n}\left(\Sigma + \underline{\mu}\underline{\mu}^T\right)^{-1}. \tag{4}$$

As a special case with the simple linear model,

$$y = \alpha + \beta x + \varepsilon, \tag{5}$$

we have

$$E(\hat{\alpha}) = \alpha, E\left(\hat{\beta}\right) = \beta, \tag{6}$$

$$var(\hat{\alpha}) \sim \frac{\sigma^2}{n}\left(1 + \frac{\mu^2}{\eta^2}\right), var\left(\hat{\beta}\right) \sim \frac{\sigma^2}{n}\frac{1}{n^2}, \tag{7}$$

where $E(x) = \mu$ and $var(x) = \eta^2$.

For the case of two covariates without an intercept,

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{8}$$

we have

$$E\left(\hat{\beta}_1\right) = \beta_1, E\left(\hat{\beta}_2\right) = \beta_2, \tag{9}$$

$$var\left(\hat{\beta}_1\right) \sim \frac{\sigma^2}{n} \times \frac{\eta_2^2 + \mu_2^2}{\left[\left(1-\rho^2\right)\eta_1^2\eta_2^2 + \left(\eta_1^2\mu_2^2 + \eta_2^2\mu_1^2 - 2\rho\eta_1\eta_2\mu_1\mu_2\right)\right]}, \tag{10}$$

$$var\left(\hat{\beta}_2\right) \sim \frac{\sigma^2}{n} \times \frac{\left(\eta_1^2 + \mu_1^2\right)}{\left[\left(1-\rho^2\right)\eta_1^2\eta_2^2 + \left(\eta_1^2\mu_2^2 + \eta_2^2\mu_1^2 - 2\rho\eta_1\eta_2\mu_1\mu_2\right)\right]} \tag{11}$$

where $E(x_1) = \mu_1$, $E(x_2) = \mu_2$, $var(x_1) = \eta_1^2$, $var(x_2) = \eta_2^2$, and $cov(x_1, x_2) = \rho\eta_1\eta_2$, with ρ being the correlation coefficient between x_1 and x_2.

Biased. If the data are generated from the model

$$y = \sum_{j=1}^{t} \beta_j x_j + \varepsilon \tag{12}$$

with $t \neq p$, we call the model (Eq. 1) biased for the data. Consider the following two biased cases:

For $t > p$, or in the case of a model underfit, we have in the large-sample limit,

$$E\left(\hat{\underline{\beta}}\right) \sim \underline{\beta} + V_X^{-1}V_{XZ}\underline{\gamma}, \tag{13}$$

$$V\left(\hat{\underline{\beta}}\right) \sim \frac{1}{n}\left[\sigma^2 + tr \times \left(V_Z - V_{ZX}V_X^{-1}V_{XZ}\right)\left(\underline{\gamma}^T\underline{\gamma}\right)\right]V_X^{-1}, \tag{14}$$

where $V_X = E(XX^T)$, $V_Z = E(ZZ^T)$, $V_{ZX} = E(ZX^T)$, and $V_{XZ} = E(XZ^T)$ with $Z = (x_{p+1}, ..., x_t)^T$ and $\underline{\gamma} = (\beta_{p+1}, ..., \beta_t)^T$.

For $t = 2$ and $p = 1$, X and Z are both one-dimensional random variables. We have

$$E\left(\hat{\beta}\right) \sim \beta + \frac{\rho\eta_x\eta_z + \mu_x\mu_z}{\eta_x^2 + \mu_x^2}\gamma, \tag{15}$$

$$var\left(\hat{\beta}\right) \sim \frac{1}{n}\left\{\frac{\sigma^2}{\eta_x^2 + \mu_x^2} + \frac{\left[\eta_z^2 + \mu_z^2 - \left(\rho\eta_x\eta_z + \mu_x\mu_z\right)^2\left(\eta_x^2 + \mu_x^2\right)^{-1}\right]\gamma^2}{\eta_x^2 + \mu_x^2}\right\}, \tag{16}$$

where $E(x) = \mu_x$, $E(z) = \mu_z$, $var(x) = \eta_x^2$, $var(z) = \eta_z^2$ and $cov(x, z) = \rho\eta_x\eta_z$.

For $t < p$, or in the case of a model overfit, we have in the large-sample limit

$$E\left(\hat{\underline{\beta}}\right) = \left(\beta_{1\times p}, 0_{1\times(t-p)}\right)^T, \tag{17}$$

$$V(\hat{\underline{\beta}}) \sim \frac{\sigma^2}{n} V_X^{-1}. \quad (18)$$

For $t = 1$ and $p = 2$, we have

$$E(\hat{\beta}_1) = \beta, E(\hat{\beta}_2) = 0. \quad (19)$$

For the variances, we have the same results as in Equations 10 and 11:

$$var(\hat{\beta}_1) \sim \frac{\sigma^2}{n} \times \frac{(\eta_2^2 + \mu_2^2)}{\left[(1-\rho^2)\eta_1^2\eta_2^2 + (\eta_1^2\mu_2^2 + \eta_1^2\mu_1^2 - 2\rho\eta_1\eta_2\mu_1\mu_2)\right]}, \quad (20)$$

$$var(\hat{\beta}_2) \sim \frac{\sigma^2}{n} \times \frac{(\eta_1^2 + \mu_1^2)}{\left[(1-\rho^2)\eta_1^2\eta_2^2 + (\eta_1^2\mu_2^2 + \eta_2^2\mu_1^2 - 2\rho\eta_1\eta_2\mu_1\mu_2)\right]}, \quad (21)$$

Misfit. If the data are generated by X_1 alone but we fit the model using X_2, then we have in the large-sample limit

$$E(\hat{\underline{\beta}}) = V_{X_2}^{-1} V_{X_2X_1} \underline{\beta_1}, \quad (22)$$

$$V(\hat{\underline{\beta}}) \sim \frac{1}{n} \times \left[\sigma^2 + tr\left(V_{X_1} - V_{X_1X_2}V_{X_2}^{-1}V_{X_2X_1}\right) \times \left(\underline{\beta_1}^T \underline{\beta_1}\right)\right] V_{X_2}^{-1}, \quad (23)$$

where $V_{X_1} = E\left(X_1X_1^T\right)$, $V_{X_2} = E\left(X_2X_2^T\right)$, $V_{X_1X_2} = E\left(X_1X_2^T\right)$, and $V_{X_2X_1} = E\left(X_2X_1^T\right)$.

If X_1 and X_2 are one-dimensional variables, we have, again in the large-sample limit,

$$E(\hat{\beta}_2) = \frac{\rho\eta_1\eta_2 + \mu_1\mu_2}{\eta_2^2 + \mu_2^2}, \quad (24)$$

$$var(\hat{\beta}_2) \sim \frac{1}{n} \times \left\{ \frac{\sigma^2}{\eta_2^2 + \mu_2^2} + \frac{\left[\eta_1^2 + \mu_1^2 - (\rho\eta_2\eta_1 + \mu_2\mu_1)^2\left(\eta_2^2 + \mu_2^2\right)^{-1}\right]\beta_1^2}{\eta_2^2 + \mu_2^2} \right\}, \quad (25)$$

where $E(x_1) = \mu_1$, $E(x_2) = \mu_2$, $var(x_1) = \eta_1^2$, $var(x_2) = \eta_2^2$, and $cov(x_1, x_2) = \rho\eta_1\eta_2$.

The above results show a complicated relation between the estimated coefficients, together with their variances, of a biased linear model and the true coefficients and covariance matrices of the explanatory variables. It would be rather hopeless to find a set of analytical expressions for the comparable situations in a GLM.

Protocol for the Construction of Synthetic Data Sets

To keep the scope of work manageable in the simulations, we considered no more than two covariates. For example, the two covariates, x_1 and x_2, could correspond to PM_{10} and CO, respectively. The dependent variable, y, could be considered as the daily mortality. In generating the synthetic data, we assumed an exact, causal, log-linear relationship between y and x_i, with the coefficients for x_i being $\beta_1 = 0.0005$ and $\beta_2 = 0.005$. These are hypothetical values used for illustrative purposes only. If we assumed the units of micrograms per cubic meter and parts per million by volume (ppmV), respectively, for x_1 and x_2, then $\beta_1 = 0.0005$ corresponds to a relative risk of 1.05 or a mortality increase of 5% per 100 μg/m^3 increase in PM_{10}, whereas $\beta_2 = 0.005$ corresponds to a relative risk of 1.005 or a mortality increase of 0.5% per ppmV increase in CO. We further assumed the intercept, α, of the exact log-linear model to be 3.132. This value corresponds to an average daily mortality of about 23.

Both x_1 and x_2 were assumed to follow a bivariate lognormal distribution with means and standard deviations (SDs) extracted from the logarithmically transformed PM_{10} and CO data for Pittsburgh during 1989–1991. In the logarithmic space, the corresponding means were 3.5 and 0.87 and the corresponding SDs were 0.619 and 0.475 for x_1 and x_2, respectively. In the concentration space, the above information essentially recovers the observed means of 40.22 μg/m^3 and 2.68 ppmV, and the observed SDs of 26.25 μg/m^3 and 1.41 ppmV, for PM_{10} and CO, respectively. In the generation of the synthetic data sets, the SD of x_1 in the logarithmic space, denoted η_1, was held fixed at 0.6 while that of x_2, denoted η_2, was allowed to vary from 0.2 to 1.0. With both η_1 and η_2 being ≤1, they are roughly proportional to the corresponding SDs in the concentration space. Note that because of the logarithmic transformation, the ηs here are not identical to the ηs described for OLMs. As a measure of confounding between the two variates, the correlation coefficient, ρ, between the variates in the logarithmic space was also allowed to vary. Again, because of the small (but realistic) values chosen for both η_1 and η_2, the correlation coefficients between the two variates in the concentration space are typically no more than 10% less than ρ for ρ = 0.5 and 0.9 and are essentially 0 for ρ = 0. Because the correlation coefficients between any two explanatory variables in Table 1 are generally positive, only positive ρs were considered in our simulations. For each realization, the values of x_1 and x_2 in the logarithmic space were generated using an S-Plus random number generator (MathSoft, Seattle, WA) for a bivariate normal distribution on an IBM RS6000 mainframe. The antilogarithms of these values were used to determine the value, m, of an exact model, $\log(m) = \alpha + \beta_1x_1 + \beta_2x_2$. In fact, m serves as the mean of the daily mortality. With this mean, the Poisson variate, y, was generated using the S-Plus random number generator for the Poisson distribution. A collection of y values with a sample size, n, constitutes the synthetic data set to be used for Poisson regression:

$$\log[E(y)] = \alpha + \beta_1x_1 + \beta_2x_2 \quad (26)$$

The sample size was also allowed to vary from 365 to 7 × 365, corresponding to a period of 1–7 years. To assure that the results of the Poisson regressions were stable, the procedure for each synthetic data set generation and the subsequent regression was performed for a total of 100 times. The means of the 100 repetitions are reported in "Results." No significant differences were found between the means with 100 repetitions and those with 1,000 repetitions.

In the Poisson regression study, the unbiased regression model contained both x_1 and x_2, as in the exact model. Several biased regression models were considered. For the case of model underfit, the synthetic data sets were constructed using the exact model containing both x_1 and x_2; the regression model assumed only x_1 as the covariate. For model overfit, the synthetic data sets were constructed using only x_1, whereas the regression model assumed both x_1 and x_2 to be the covariates. For model misfit, two cases were considered. First, the synthetic data sets were based on only x_1; the regression model contained only x_2 as the covariate. Second, only x_2 was used in the synthetic data sets; only x_1 was the covariate in the regression. The latter is not equivalent to the former because we always allowed only the range of x_2 to vary.

Results

The impact of 1) confounding or correlation, ρ, between x_1 and x_2; 2) the data range or SD, η_2, of x_2; and 3) the sample size, n, on the outcome of the Poisson regression will be presented in the same order as described for OLMs. The outcome is represented by the estimates of the coefficients, $E(\hat{\beta}_i)$, and their respective t-values. A t-value ≥2 is considered significant. In all cases, the

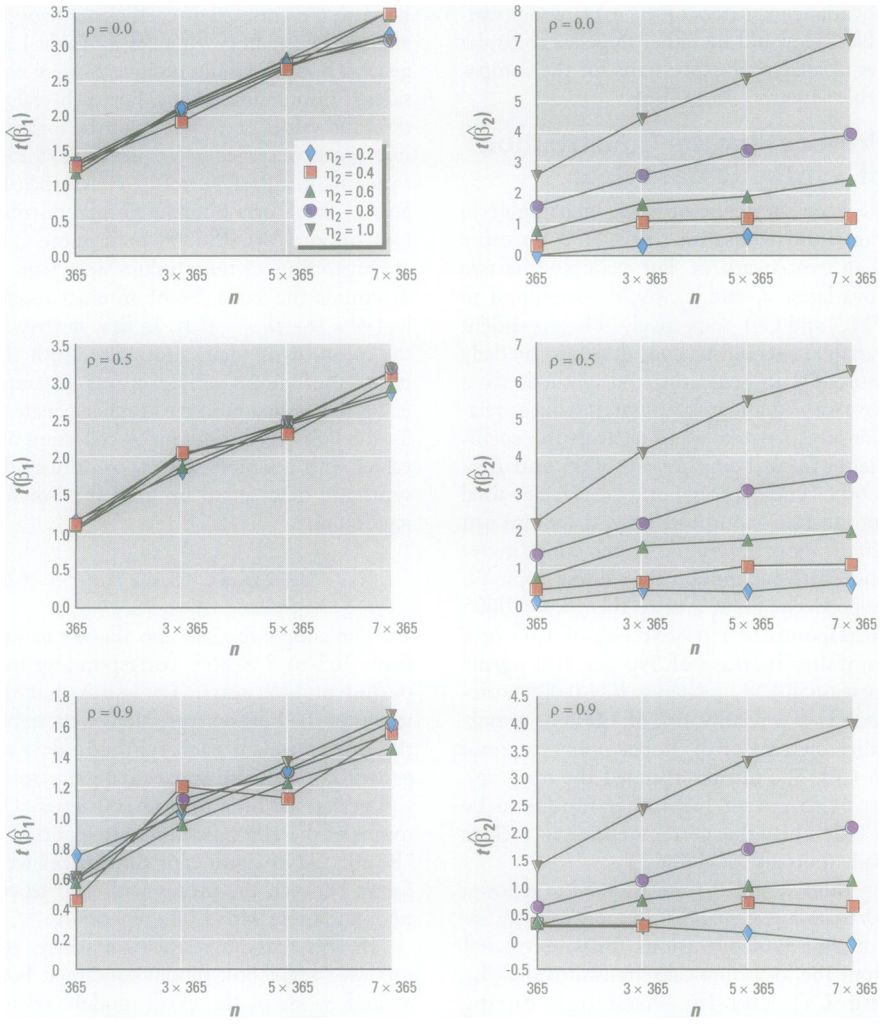

Figure 1. Behavior of $t(\hat{\beta}_1)$ and $t(\hat{\beta}_2)$ as a function of n (number of observations), η_2 (standard deviation of log x_2, representing the range of the x_2 covariate) and ρ (correlation coefficient between log x_1 and log x_2 and comparable to that between x_1 and x_2) for an unbiased Poisson regression using the same covariates as the exact model.

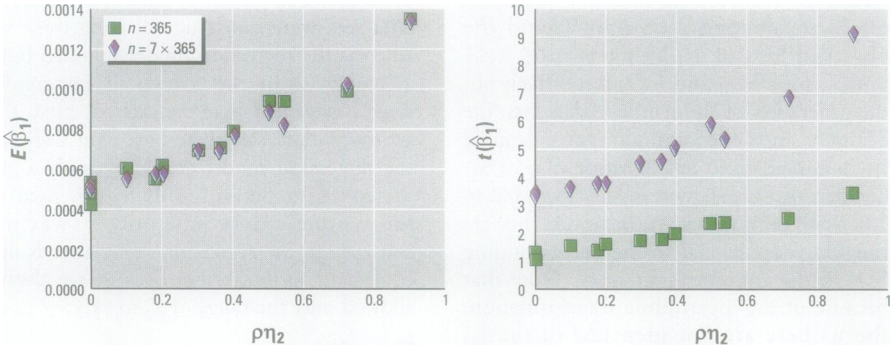

Figure 2. Behavior of $E(\hat{\beta}_1)$ and $t(\hat{\beta}_1)$ as a function of $\rho\eta_2$, and n for an underfitting Poisson regression containing x_1 as the covariate to describe data created from an exact model containing x_1 and x_2.

overdispersion parameter, estimated as the residual deviance divided by the model degrees of freedom, is within 1% of 1.

Unbiased. Neither $E(\hat{\beta}_1)$ nor $E(\hat{\beta}_2)$ are impacted by the correlation between the two covariates. The estimates are not significantly different from the coefficients of the exact model. This is consistent with Equation 9 for the OLMs. However, both $t(\hat{\beta}_1)$ and $t(\hat{\beta}_2)$ decrease with increasing ρ. This is qualitatively consistent with the increasing variance of the estimated coefficients when ρ increases in Equations 10 and 11 for the OLMs. Also, for the parameters used in the exact model, Equation 10 indicates a weak dependence of $t(\hat{\beta}_1)$ on η_2, whereas Equation 11 indicates an essentially linear dependence of $t(\hat{\beta}_2)$ on η_2. Figure 1 shows both $t(\hat{\beta}_1)$ and $t(\hat{\beta}_2)$ as a function of n, ρ, and η_2. From Figure 1, one sees that the dependence of $t(\hat{\beta}_2)$ on η_2 is not linear.

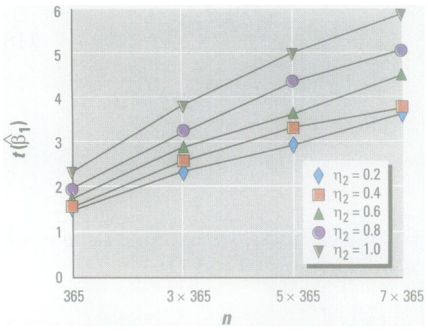

Figure 3. Behavior of $t(\hat{\beta}_1)$ as a function of n and η_2, given ρ = 0.5, for an underfitting Poisson regression containing x_1 as the covariate to describe data created from an exact model containing x_1 and x_2.

There is no major qualitative difference in the behavior of the estimated coefficients between the OLM and the Poisson regression when all causal variables are covariates in the regression model. An unbiased regression model can recover the estimated coefficients of the covariates. But if the correlation among the covariates increases, the significance of the estimated coefficients decreases.

Biased. In the underfit case, the synthetic data sets contain the effect of both x_1 and x_2, whereas the regression model contains only x_1 as the covariate. In the ordinary linear regression, Equation 15 indicates that $E(\hat{\beta}_1)$ increases with ρ and η_2, or more precisely, $E(\hat{\beta}_1)$ is asymptotically β_1 plus a term that is linearly related to $\rho\eta_2$. This is qualitatively consistent with a nearly linear relation observed in the Poisson regression (Fig. 2). The estimated coefficient departs significantly from the β_1 of the exact model as ρ and η_2 increase. The variance of $\hat{\beta}_1$ decreases with ρ in the asymptotic expression (Equation 16) for the OLM. This is again consistent with the simulation result that $t(\hat{\beta}_1)$ increases with ρ. In addition, $t(\hat{\beta}_1)$ also increases with η_2, and this increase is enhanced by an increasing ρ (Fig. 2). As the sample size, n, increases, $t(\hat{\beta}_1)$ increases as well (Figs. 2, 3). If x_2 were used in the regression model, then based on Equations 15 and 16 and the parameters of the exact model, one would expect $E(\hat{\beta}_2)$ to be essentially proportional to ρ/η_2 and $t(\hat{\beta}_2)$ to again increase with ρ.

The above result has a profound implication insofar as the PM mortality studies are concerned. Underfitting is very likely

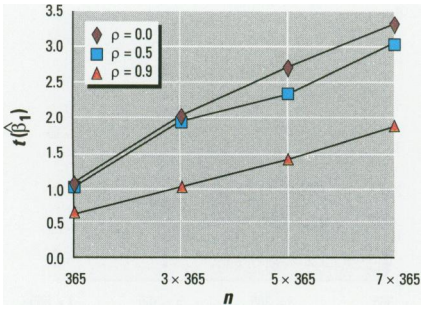

Figure 4. Behavior of $t(\hat{\beta}_1)$ as a function of n and ρ given $\eta_2 = 0.2$, for an overfitting Poisson regression containing x_1 and x_2 as the covariates to describe data created from an exact model containing x_1.

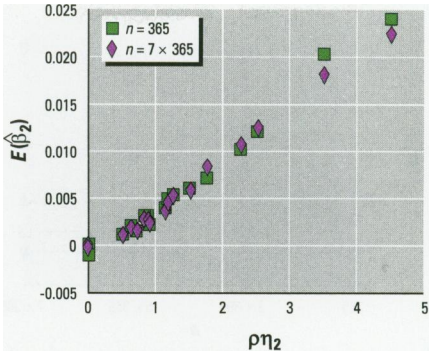

Figure 5. Behavior of $E(\hat{\beta}_2)$ as a function of n and ρ/η_2 for a misfitting Poisson regression containing x_2 as the covariate to describe data created from an exact model containing x_1.

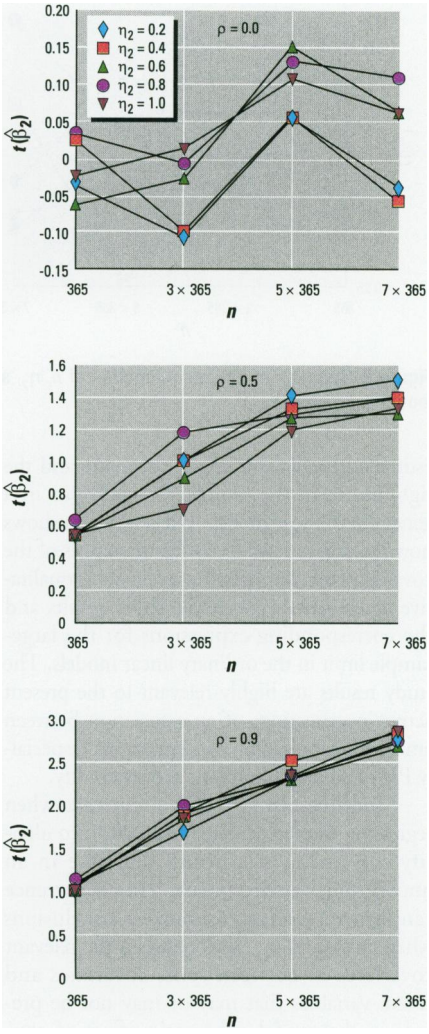

Figure 6. Behavior of $t(\hat{\beta}_2)$ as a function of n, η_2, and ρ for a misfitting Poisson regression containing x_2 as the covariate to describe data created from an exact model containing x_1.

when the number of covariates used is small (e.g., one or two). If a causal variable such as CO is missing in the regression model and the variable is highly correlated with a covariate (e.g., PM) in the regression model, then the regression model will indicate a strong but erroneous association of the dependent variable or effect (the daily mortality, for example) with the covariate. In fact, the estimated coefficient of the covariate will be compromised by the size of the actual coefficient of the missing variable, the range of the missing variable, as well as the magnitude of the correlation coefficient between the covariate and the missing variable. The *t*-value of the estimated coefficient also increases with the correlation coefficient and the range of the missing variable. In addition, increasing the sample size (to several years of data, for example) also increases the *t*-value, actually making the erroneous association appear more convincing.

In the overfit case, the synthetic data sets were constructed using only x_1; the regression model contains both x_1 and x_2 as the covariates. In agreement with Equation 19 for the ordinary linear regression, the estimated coefficients of both covariates are not significantly different from their exact values, being zero for $E(\hat{\beta}_2)$. They are not affected by the correlation coefficient between the two covariates. The $t(\hat{\beta}_1)$, on the other hand, decreases with increasing ρ, and does not depend strongly on η_2. Figure 4 shows $t(\hat{\beta}_1)$ as a function of ρ and n, with η_2 held constant at 0.2. As expected, $t(\hat{\beta}_2)$ is essentially zero. If the exact model contains only x_2, one expects $t(\hat{\beta}_1)$ to be zero and $t(\hat{\beta}_2)$ to be increasing with η_2 and decreasing with ρ.

The above result indicates that overfitting should not lead to a serious bias in the estimated coefficients of the covariates, but the correlation between the causal and redundant covariates will reduce the significance of the estimated covariates.

Misfit. In the first misfit case, x_1 was used in the exact model and x_2 was the covariate in the regression model. Even though x_2 plays no role in the dependent variable of the synthetic data sets, in the regression model x_2 influences $E(\hat{\beta}_2)$ and $t(\hat{\beta}_2)$ through ρ. For the ordinary linear regression, Equation 24 shows that $E(\hat{\beta}_2)$ increases with increasing ρ and decreases with increasing η_2. Figure 5 shows $E(\hat{\beta}_2)$ as a function of ρ/η_2. The significance of the estimate, $t(\hat{\beta}_2)$, increases with ρ and n, but, interestingly, not with η_2 (Fig. 6).

In the second misfit case, x_2 was used in the exact model and x_1 was the covariate in the regression model. In this case, the variation in η_2 directly impacts the dependent variable in the synthetic data sets. Figure 7 shows $E(\hat{\beta}_1)$ and $t(\hat{\beta}_1)$ as an increasing function of $\rho\eta_2$. Again, no causal meaning can be attached to the magnitude of the estimated coefficient. Even so, Figure 8 shows that $t(\hat{\beta}_1)$ can be large and increasing with ρ and η_2, in contrast with $t(\hat{\beta}_2)$ above, which has little or no dependence on η_2.

Model misfit is again a likely occurrence. The result of the misfit is a set of totally meaningless estimated coefficients, yet with increasing significance as the sample size and the ranges of the missing causal variables increase and as the correlation between the covariates and the true causal variables increases. The potential for misleading inference in model misfit in epidemiological studies cannot be overemphasized.

Conclusion

Using synthetic data sets, the present study shows the impact of confounding on the

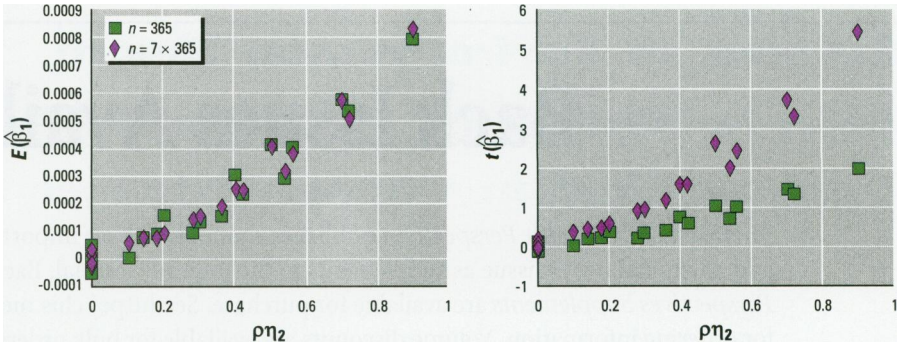

Figure 7. Behavior of $E(\hat{\beta}_1)$ and $t(\hat{\beta}_1)$ as a function of $\rho\eta_2$ and n for a misfitting Poisson regression containing x_1 as the covariate to describe data created from an exact model containing x_2.

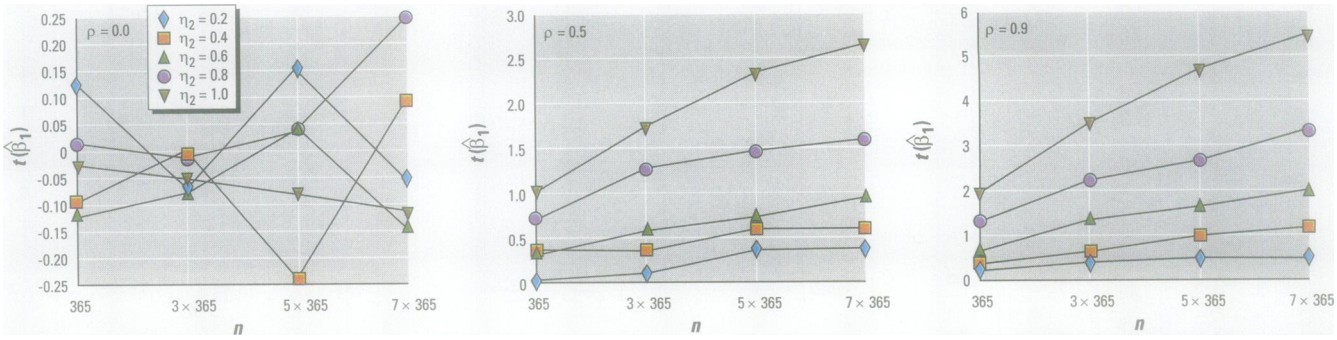

Figure 8. Behavior of $t(\hat{\beta}_1)$ as a function of n, η_2, and ρ for a misfitting Poisson regression containing x_1 as the covariate to describe data created from an exact model containing x_2.

estimated coefficients of the covariates and the significance of the estimated coefficients in a generalized linear model. The study also shows how this impact changes with the ranges of the covariates and the sample size. There is qualitative agreement between the study results and the corresponding expressions for the large-sample limit in the ordinary linear models. The study results are highly relevant to the present active investigations of an association between ambient air pollutant concentrations (especially PM) and daily mortality and morbidity.

Modeling bias is a likely occurrence when regression models are used in an effort to identify the causes of a health outcome in an uncontrolled environment. This occurrence can lead to seriously erroneous conclusions when confounding exists between the relevant covariates or between some covariates and causal variables that may or may not be present in the model. The main effect of confounding for model overfit is a reduction in the significance of the estimated coefficients. The effect of model underfit or misfit (a more common occurrence), on the other hand, leads not only to erroneous estimated coefficients, but a misguided confidence, represented by large t-values, that the estimated coefficients are significant. The results of this study indicate that models that use only one or two air quality variables such as PM_{10} and SO_2 are likely unreliable, and that models containing several correlated and toxic or potentially toxic air quality variables should also be investigated in order to minimize the situation of model underfit or misfit. It is also possible that models containing more pollutants as covariates may have estimated coefficients that are unphysical or counter-intuitive. Such a situation would call for controlled experiments to establish a causal relation between a pollutant or multiple pollutants and a health end point.

References and Notes

1. National ambient air quality standards for particulate matter. Fed Reg 62(138):38651–38701 (1997).
2. Vedal S. Ambient particulates and health: lines that divide. J Air Waste Manage Assoc 47:551–581 (1997).
3. Schwartz J, Dockery DW. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am Rev Respir Dis 145:600–604 (1992).
4. Moolgavkar SH, Luebeck EG, Hall TA, Anderson EL. Air pollution and daily mortality in Philadelphia. Epidemiology 6:476–484 (1995).
5. Samet JM, Zeger S, Berhane K. The association of mortality and particulate air pollution. In: Particulate Air Pollution and Daily Mortality: Replication and Validation of Selected Studies. The Phase I Report of the Particle Epidemiology Evaluation Project. Cambridge, MA:Health Effects Institute, 1995;1–122.
6. Moolgavkar SH, Luebeck EG. A critical review of the evidence on particulate air pollution and mortality. Epidemiology 7:420–428 (1996).
7. Samet JM, Zeger SL, Kelsall JE, Xu J, Kalkstein LS. Air pollution, weather, and mortality in Philadelphia 1973–1988. In: Particulate Air Pollution and Daily Mortality: Analyses of the Effects of Weather and Multiple Air Pollutants. The Phase I.B Report of the Particle Epidemiology Evaluation Project. Cambridge, MA:Health Effects Institute, 1997;1–44.
8. Schwartz J, Dockery DW. Particulate air pollution and daily mortality in Steubenville, Ohio. Am J Epidemiol 135:12–19 (1992).
9. Schwartz J. Air pollution and daily mortality in Birmingham, Alabama. Am J Epidemiol 137:1136–1147 (1993).
10. Davis JM, Sacks J, Saltzman N, Smith R, Styer P. Airborne Particulate Matter and Daily Mortality in Birmingham, Alabama. Tech Report No. 55. Research Triangle Park, NC:National Institute of Statistical Sciences, 1996.
11. Moolgavkar SH, Luebeck EG, Hall TA, Anderson EL. Particulate air pollution, sulfur dioxide, and daily mortality: a reanalysis of the Steubenville data. Inhal Toxicol 7:35–44 (1995).
12. Styer P, McMillan N, Gao F, Davis J, Sacks J. The effect of outdoor airborne particulate matter on daily death counts. Environ Health Perspect 103:490–497 (1995).
13. Lipfert FW, Wyzga RE. Air pollution and mortality: issues and uncertainties. J Air Waste Manage Assoc 45:949–966 (1995).
14. Smith RL, Davis JM, Speckman P. Human health effects of environmental pollution in the atmosphere. In: Statistics for the Environment 4: Health Effects. (Barnett V, Turkman F, Stein A, eds). Chichester:John Wiley. In press.
15. Seber GAF. Departures from Underlying Assumptions. In: Linear Regression Analysis. New York:John Wiley & Sons, 1977;140–176.
16. Chen C, Zhang YG. Large Sample Properties of A Series Estimations of Cross-Validation in Linear Regression Models. Tech Rpt No 96-19. West Lafayette, IN:Purdue University, Department of Statistics, 1996.
17. McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. London:Chapman & Hall, 1989.